Chapter 1 Solutions

Sidhanth Holalkere

January 24, 2021

1. *Self-Play* Suppose, instead of playing against a random opponent, the reinforcement learning algorithm described above played against itself, with both sides learning. What do you think would happen in this case? Would it learn a different policy for selecting moves?

   - I think that instead of learning to play against just the opponents (which presumably are humans), it would learn a more general and probably more optimal form of play. Instead of selecting the moves that work well against random opponents (who are most likely playing flawed) it would have to make more generally optimal moves. We see this in the real world through work done by OpenAI who have trained *tabula rasa* models to a superior level than previously created.

2. *Symmetries* Many tic-tac-toe positions appear different but are really the same because of symmetries. How might we amend the learning process described above to take advantage of this? In what ways would this change improve the learning process? Now think again. Suppose the opponent did not take advantage of symmetries. In that case, should we? Is it true, then, that symmetrically equivalent positions should necessarily have the same value?

   - We can amend the learning process by taking advantage of the 4-sided rotational symmetry of a tic-tac-toe board and quarter the sizes of the action and state space. This would probably make the learning process faster because it would only have to learn for a quarter of the number of states. If the opponent did not take advantage of symmetries then we should not as well. If for two symmetrically equivalent positions, if the opponent plays a good move for one and a bad move for another we wouldn't be able to exploit the opponents mistake.

3. *Greedy Play* Suppose the reinforcement learning player was *greedy*, that is, it always played the move that brought it to the position that it rated the best. Might it learn to play better, or worse, than a noongreedy player? What problems might occur?

   - Assuming the player has no prior information about the state or state-action values, the greedy player will learn to play worse than a nongreedy player. This is because once the greedy player identifies one move at a certain state that it has rated any bit higher than the others, it will only choose that move from that on (in that state) and fail to explore whether other moves are better.

4. *Learning from Exploration* Suppose learning updates occured after all moves, including exploratory moves. If the step-size parameter is appropriately reduced over time (but not the tendency to explore), then the state values would converge to a different set of probabilities. What (conceptually) are the two sets of probabilities computed when we do, and when we do not, learn from exploratory moves? Assuming that we do doncinue to make exploratory moves, which set of probabilities might be better to learn? Which would result in more wins?

   - When we do learn from exploratory moves, certain state values will be lower than they should be (due to exploring a "bad" move and then lowering the previous state value). This does not happen if we do not learn from exploratory moves. It is better to not learn from exploratory moves so we actually get closer to the true state values and result in more wins.

5. *Other Improvements* Can you think of other ways to improve the reinforcement learning player? Can you think of any better way to solve the tic-tac-toe problem as posed?

- Since tic-tac-toe is a relatively simple game and every possible trajectory of moves is easily calculable and storeable, one way we could improve the reinforcement learning player would be to initialize the state values to be related to the probability of winning (when considering all possible combinations of next moves). Then it would start out with a good baseline for which moves tend to lead to better outcomes, while still being able to explore whether the opponent blunders in certain states.